EduCam: Cinematic Vocabulary for Educational Videos

Soja-Marie Morgens and Arnav Jhala
smorgens@soe.ucsc.edu and jhala@soe.ucsc.edu
Computer Science Department, U. of California, Santa Cruz, CA

Abstract
Cinematography relies on a rich vocabulary of shots and framing parameters to communicate narrative event structures. This cinematic vocabulary stems from the physical capabilities of camera movement and uses aesthetic aspects to engage the audience and influence audience comprehension. At present, automatic physical camera control is limited in cinematic vocabulary. Virtual cameras, on the other hand, are capable of executing a variety of cinematic idioms. Specifically, shot composition, match cut, and establishing shot sequence are three common cinematic techniques often lacking in automated physical camera systems. This paper proposes an architecture called EduCam as a platform for employing virtual camera control techniques in the limited stochastic physical environment of lecture capturing. EduCam uses reactive algorithms for real time camera coverage with hierarchical constraint satisfaction techniques for offline editing.

This paper focuses on three cinematic idioms that are common in professionally-shot educational videos and documentaries: shot composition, match cut, and the establishing shot sequence. It proposes EduCam, an architecture that allows for the employment of virtual camera control algorithms in the physical lecture capture environment. EduCam uses real time reactive camera coverage to allow for offline editing with hierarchical constraint satisfaction algorithms. This paper will evaluate the current physical automatic camera control for their cinematic capacity, discuss current virtual cinematography and finally outline how EduCam allows for hierarchal constraint satisfaction editing in a physical world.

Amateur video has become an important medium for disseminating educational content. The DIY community routinely creates instructional videos and tutorials while professional educators publish entire courses on Gooru and Coursera (Coursera 2014; Gooru 2014). Video gives educators the ability to flip the classroom, pushing passive viewing outside of the classroom and bringing interactivity and discussion back in. With video, educators have access to cinematic idioms to shape the narrative of their material.

However these idioms rely on a high threshold of videography knowledge, equipment, and time. So educators use static cameras with time-consuming editing or automatic lecture capture systems. Current automatic lecture capture are passive capture systems, created to record the oratory style of the educator to his live audience rather than employ cinematic vocabulary. Moreover, their coverage typically consists of just a few cameras in the back of the classroom. However this reduces their shot vocabulary to long and medium shots with limited shot composition capability. Current virtual camera control algorithms are able to employ the full variety of shots and some cinematic transitions. However they do stochastic physical environment and typically assume unlimited camera coverage.

Figure 1: An example of good shot composition from Maltese Falcon. The subject is on the upper left third lines and his gaze is into the open space. (Bordwell 2014)

Cinematic Vocabulary
The three most common cinematic idioms used to engage audiences in content-driven film are shot composition, match cuts, and establishing shot sequence.

Shot Composition ensures the shot is visually engaging and cues the audience to the subject of the shot. Generally a good engaging composition employs rule of thirds and is from an angle that emphasizes perspective. Subjects are typically on the side of the frame opposite of where they are moving or gazing (see Figure 1).

Match Cuts are used to smoothly transition an audience’s eyes from one shot to another. In it the area of interest from the second shot is often matched with the area of interest in the first shot as shown in Figure 2. This maintains the audience’s eyes on the subject and prevents “jitteriness” or loss

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
of focus as the audience re-engages their eyes with the subject.

Establishing shot sequence is used to signal a change of scene or subject. It consists of sequential shots start with a long shot then transition to medium shot and finally end on close shots of the subject. Used as the beginning of a film, it focuses the audience’s attention on the subject and eases the audience into the story. Used as a chapter break, it signals to the audience a change and allows them time to refocus.

These three methods rely on the director’s ability to manipulate multiple cameras along seven degrees of freedom according to the subject’s position and future intentions. In computational cinematography the virtual world allows perfect knowledge of a subject’s location and near perfect knowledge of the subject’s intentions. Coverage is not an issue and most systems rely on multiple cameras. However in lecture recording systems, knowledge of future intent is imperfect; editing algorithms are limited to real time operations; cameras are limited in number and mobility. By using real time capture from multiple angles and zooms but non real time editing, the EduCam system will allow for use of a larger variety of cinematic idioms without loss of quality.

Lecture Capture Domain

The lecture capture domain is an excellent testbed for bringing virtual cinematographic techniques into the real world. While the educational video exists within physically stochastic environment it is one filled with discernible cues and recognizable subjects. It has just two subjects: a single instructor and the material which is either written on the board or shown as a live demo. Movement is long a bounded plane and at a limited velocity, allowing for tracking cameras to be set in pre-determined positions without significant loss of coverage. Facial recognition and motion extraction algorithms allow cameras to roughly track the subjects in real time with finessed shot composition done offline. Cues to transition between the subjects can be taken from the narrator’s gaze, gesture and rhythm.

Physical Camera Control Systems

Automatic physical camera control systems can be divided up into two categories: passive and active systems. Passive systems are unobtrusive and record the narrator’s oration to a live audience, whether in a meeting or a classroom. Active systems are set up as the primary medium through which the narrator communicates.

Current lecture capture systems are passive and use real-time editing. These systems typically consist of one or more static PTZ (Pan, Tilt, Zoom) cameras. Kushalnagar, Cavender, and Pâris involves a single camera in the back of the classroom to capture the narrator and the material in a long shot. More complex set ups such as Liu et al.2001 and Lampi, Kopf, and Effelsberg involve a camera on audience as well as a camera that can actively track the narrator. Lampi, Kopf, and Effelsberg does use some cinematic vocabulary: it composes its shots according to rule of thirds in its real time coverage.

These passive capture systems, limited in their coverage and real time editing constraints, employ simple auto directors without cinematic editing. Liu et al. (2001) and Lampi, Kopf, and Effelsberg transition based on shot length heuristics and physical buttons. Liu et al. (2002) and Kushalnagar, Cavender, and Pâris do not transition but show all streams simultaneously.

Active capture systems demonstrate the capabilities of current technology for reactive real time automated control and the need for cinematic idioms to improve audience engagement (Chi et al. 2013; Birnholtz, Ranjan, and Balakrishnan 2010; Ranjan et al. 2010; Ursu et al. 2013). Ranjan et al. tracks the current speaker in a meeting through visual facial recognition and audio cues. Both methods prove to be noisy. However, it reported its output as comparable to real time editing video crew and is capable of close, medium, and long shots. Birnholtz, Ranjan, and Balakrishnan uses motion tracking of a hand to determine the material subject and when to transition between shots. Its system can employ long shots and close shots and found that the use of a variety of shots shapes audience engagement. Chi et al. creates video using a single camera that follows the narrators position via Kinect recognition algorithms. It demonstrated the ease at which a Kinect can track an object adequately during live recording and the reluctance of a narrator to participate in changes of camera coverage.

Overall, these studies show that physical systems have the capability to do real time reactive camera coverage. However while active systems do use a variety of shots, they do not often use cinematic vocabulary for shot transitions. They either have no transitions or use simple state machines that transition based on shot length heuristics and audio cues.

Virtual Camera Control

Automatic virtual camera control can generally be divided into three approaches: reactive, optimization-based, and constraints satisfaction (Christie, Olivier, and Normand 2008).

The reactive approaches are adapted from robotics and best suited to real-time control of camera coverage. They look for local optima near the camera’s current position and generally focus on minimizing occlusion of the subject and optimizing shot composition. For example, Ozaki et al. uses A* with a viewpoint entropy evaluation based on occlusion and camera collision with the environment.

Optimization approaches tend to focus on shot compo-
sition but interactive editing systems provide experts with a range of temporal cinematic idioms of movement. These optimization approaches have used frame coherence and shot composition with static cameras (Halper and Olivier 2000). Interactive systems take directorial input to create local optima for use with cinematic vocabulary such as the Semantic Volumes presented in Lino et al. (2010). However both methods heavily rely on full knowledge of the environment and animation paths.

Constraint Satisfaction approaches focus on a set of constraints on camera parameters that each shot must obtain. While many of these approaches focus on shot composition related parameters, Jardillier and Languenou also includes temporal constraints, camera movement constraints, and framing to differentiate between different types of shots. However pure constraint satisfaction approaches result in a complex constraint set and an unwieldy language to specify these constraints. This makes it difficult for users to avoid over- and under-constraining their specifications. Hierarchical constraint satisfaction, as used in ConstraintCam, allows for specification of relationships between constraints that enables better specification of relaxation priorities of constraints as well the specification for possibly conflicting constraints for the constraint solver. (Bares, Grégoire, and Lester 1998).

EduCam

EduCam is a platform for employing computational cinematic control algorithms on a physical platform. This section will outline the physical elements of EduCam, its real time reactive algorithm for camera control and its hierarchical constraints satisfaction rules for editing.

Physical Camera Setup

EduCam has a set up that allows for extended camera coverage and is able to track the two subjects appropriately in real time. EduCam uses a traditional four camera set up about the action line as shown in Figure 4. The two outside cameras capture visually engaging medium and close shots. One middle camera captures long and medium shots, and the other middle camera takes close material shots. Camera placement is flexible and can be adjusted based on the needs of the video and the patterns of the narrator. By placing the cameras about the action line rather than in difference to a live audience, this system has the physical capability to engage in shot composition, match cut and the establishing shot sequence.

Like Ranjan et al., the EduCam system employs a dual camera system as seen in Figure 3. In this case, the Kinect and Raspberry Pi use a reactive coverage algorithm with facial detection and motion extraction algorithms to actively maintain the subject in the current Region of Interest (ROI). The camera on top (currently a GoPro) passively records the scene in HD for offline editing.

Real Time Reactive Camera Coverage

Reactive algorithms is used for the initial real time camera coverage. At all points, all three cameras are engaged in subject tracking through facial detection, and motion extraction algorithms standard in the OpenCV and OpenNI libraries. At each time point however, they maintain the subject at the same ROI. This will allow there to be a match cut at any time. The three cameras will also keep the shot as a medium shot. This allows not only for match cut at any time period but for shot composition to be further refined as necessary after shooting. Many amateur videographers attempt to capture their material cropped correctly the first time, leaving them with little edit if they see a mistake. Using 1920x1080 footage instead of the standard 640x480 provided by the Kinect, a medium shot can be converted to a close shot using digital zoom.
Hierachical Constraint Satisfaction

EduCam proposes to use hierachical constraint satisfaction with weighted rules in the post-editing process to transition between shots. The hierachical nature of content in educational videos could naturally fit the hierachical specification language for constraints. Each shot is optimized for its shot composition and given a score on composition, best meeting the 3 types of shots, and creating a match for the current shot. Cues to transition include current shot length, audio pauses to signal scene change, or simply a minimized error between the matching of two shots. Rule weights change temporally depending on the current cues and the next shot is chosen that best meets constraints. The current constraint set for EduCam includes rules for match cuts, establishing shots, screen time for narrator and content, and shot composition. EduCam will be evaluated based on its ability to engage a viewing audience with the material by looking for correlations between engagement related observations and constraints.

Conclusion

The technology to run physical automated camera control with cinematic vocabulary is available. Both the active and passive lecture capture systems have demonstrated that narrator tracking can be done in real time and at low cost. Computational cinematic algorithms has demonstrated that autonomous cinematography works in limited stochastic environments. Moreover the educational video domain is an ideal transition point from virtual to the physical domain. Once EduCam is fully implemented, there are several next steps. The first and foremost is to evaluate whether these cinematic idioms, once employed in education, will not only engage students in the material better but also engage educators in creating better material. Secondly, the EduCam set up can be applied to a more complex physical environment, such as a play or even a meeting, where multiple subjects must be tracked. Dividing the automatic director control into rough reactive and constraint satisfaction allows for such developments. EduCam is the bridge towards automatic cinematic control in a real stochastic environment.

References


